





Semestre : 2
Module : Méthodes Quantitatives
Elément : Statistiques
Enseignant : Mr BENMOUSSA

Eléments du cours

- Définitions et éléments du vocabulaire
- Réalisation d'un tableau statistique
- Graphiques statistiques de base
- Les caractéristiques de tendance centrale et de position
- Les moyennes
- Les paramètres de dispersion
- La concentration
- L'étude de la corrélation

Numérisation & Conception
Mr Mohamed-Fadil ZIADI

Chapitre 1 : DEFINITIONS ET ELEMENTS DU VOCABULAIRE

I- Définitions et objets :

La statistique est la science qui a pour objet de recueillir un ensemble de données numériques relatives à tel ou tel phénomène aléatoire et d'exploiter cette information pour établir toutes relations de causalité par l'analyse et l'interprétation.
Un phénomène aléatoire est un phénomène comportant des variables aléatoires, c'est à dire des variables liées au hasard et dont les valeurs ne peuvent en conséquence être connues d'avance. (Exemple : le nombre de points marqué par un dé)

On distingue :

- La statistique descriptive ou statistique de constatation, qui concerne les tableaux, les graphiques relatives à des inventaires, des enregistrements, des recensements…etc.
- La méthode statistique qui concerne l'ensemble des procédés et méthodes pour l'analyse et l'interprétation.

1- Domaines d'application :

Le domaine d'utilisation de la statistique est tellement étendu qu'il serait impossible de citer toutes les applications, mais on va citer quelques exemples :
- La recherche biologique et médicale ;
- La recherche spatiale ;
- Le contrôle des fabrications dans l'industrie ;
- Le sondage d'opinion ; les enquêtes de marché ;
- Les assurances ;
- Les recherches opérationnelles ;
- L'étude de la conjoncture ;
- La détermination des indices économiques.

Les sondages d'opinion, en particulier, ont connu une extension considérable au cours de ces dernières années.
Les perfectionnements considérables intervenus dans le domaine des machines à calculer ont contribué à étendre les possibilités de la statistique.

2- Ensemble, sous-ensembles, unités :

☛ Un ensemble ou référentiel statistique composé d'éléments ou d'unités statistiques est dit population ou univers.
☛ Un sous-ensemble de l'ensemble est un échantillon.
☛ Une unité statistique doit être définie sans ambiguïté. Elle peut comporter de nombreux caractères, ceux-ci pourront eux même comporter plusieurs modalités.



Exemple :
* Ensemble de production : 135 ouvriers de l'usine X ;
* Echantillon : 5 ouvriers ;
* Unité : 1 ouvrier de l'usine X ;
* Caractères : a- le salaire (modalités)
b- ancienneté (modalités)

3- Caractères qualitatifs et quantitatifs continus ou directs :

☛ Les caractères quantitatifs sont ceux auxquels on peut attribuer une valeur numérique. (Exemple : une taille)
☛ Les caractères qualitatifs sont ceux auxquels on peut seulement associer une valeur numérique arbitraire. (Exemple : une couleur)
☛Un ensemble coordonné des valeurs d'un caractère quantitatif constitue une suite ou série statistique.
☛ Un caractère continu est un caractère qui peut prendre n'importe quelle valeur numérique. (Exemple : une surface, un prix)
☛ Un caractère discret ou discontinu est un caractère qui ne peut prendre que des valeurs isolées en général des nombres entiers. (Exemple : nombre de personnes dans une famille). Dans le cas d'un caractère discontinu, l'interprétation est dénuée de sens.

II- Elaboration des statistiques :

- La collection des renseignements ;
- Le recensement et sondage ;
- Enquêtes et questionnaires ;
- Dépouillement.

Chapitre 2 : REALISATION D'UN TABLEAU STATISTIQUE

Un tableau permet une présentation synthétique des informations recueillies, il doit se suffire à lui même, c'est pourquoi il est nécessaire qu'il comporte les indications suivantes :
* Le titre indiquant l'objet du travail statistique ;
* L'unité de mesure qui a été utilisée ;
* La référence de la source de la documentation.

I- Présentation d'un tableau :

D'une façon générale, un tableau se compose :
* D'une colonne indiquant les différentes modalités de la variable x_i ;
* D'une ou plusieurs autres colonnes indiquant l'effectif correspondant à ces diverses modalités.
Mais selon que la variable est discrète ou continue, les tableaux se présentent de la façon suivante.

1- Tableau concernant une variable directe :

☛ **Exemple :**

Distribution du personnel d'une entreprise en fonction du nombre d'enfants.

Nombre d'enfants x_i	Effectifs n_i
0	12
1	31
2	29
3	11
4	4
5	2
6 et +	1
	Total : 90

☛ **Remarque : Lecture du tableau :**

12 membres du personnel portent zéro enfant.
31 membres du personnel portent un enfant.

2- Tableau concernant une variable continue :

☛ **Exemple :**

Distribution des réceptions de marchandises en fonction du nombre de colis.

Nombre de colis x_i	Effectifs n_i
1 à 5	20
6 à 10	30
11 à 15	60
16 à 20	50
21 à 30	30
31 et +	10
	Total: 200

☛ **Remarque : lecture du tableau :**

* Nous avons reçu 20 fois des livraisons contenant des colis entre 1 et 5.
* Les amplitudes peuvent être inégales.
* La colonne des x_i peut être présentée de la façon suivante :
1 – 5
5 – 10
10 – 15
↓ etc.

Mais il faut indiquer clairement la convention de bornage choisi.

II- Notions de fréquence :

La deuxième colonne d'un tableau de statistique enregistre le nombre de fois que la valeur de la variable mentionnée dans la première colonne a été rencontrée.
Il s'agit d'une fréquence, notée f_i, et celle-ci peut apparaître sous divers aspects selon les critères retenus.

1- Fréquence absolue, fréquence relative :

* La fréquence absolue, comme son nom l'indique, donne le nombre d'unités en valeurs absolues.
* La fréquence relative est calculée en divisant chaque fréquence absolue par l'effectif total de la population.
En d'autres termes, la fréquence est exprimée en valeurs relatives multipliée par 100 donne un pourcentage.

Valeur de colis x_i	fréquences f		
	absolues	relatives	
1 à 5	20	0,1	10 %
6 à 10	30	0,15	15 %
11 à 15	60	0,3	30 %
16 à 20	50	0,25	25 %
21 à 30	30	0,15	15 %
31 et +	10	0,05	5 %
	total: 200	total: 1	total: 100 %

☛ **Remarques :**

* Pour les fréquences relatives, le tableau se lis comme suit :
  10 % des livraisons reçues contenaient entre 1 et 5 colis.
* La somme des fréquences relatives est toujours égale à 1.

2- Fréquences simples, fréquences cumulées :

* Les fréquences simples, qu'elles soient absolues ou relatives, indiquent comment se distribue la variable par rapport aux différentes modalités.
* Les fréquences cumulées, qu'elles soient absolues ou relatives, indiquent comment se répartis la variable par rapport aux différentes modalités. Il existe deux catégories de fréquences cumulées :

- Les fréquences cumulées croissantes qui indiquent combien d'unités de la population sont caractérisées par une valeur inférieure ;
- Les fréquences cumulées décroissantes qui indiquent combien d'unités de la population sont caractérisées par une valeur supérieure.

Valeur de colis. x_i	Fréquences absolues			Fréquences relatives		
	Simples	Cumulées		Simples	Cumulées	
		Croissantes	Décroissantes		Croissantes	Décroissantes
1 à 5	20	20	200	0,1	0,1	1
6 à 10	30	50	180	0,15	0,25	0,9
11 à 15	60	110	150	0,3	0,55	0,75
16 à 20	50	160	90	0,25	0,8	0,45
21 à 30	30	190	40	0,15	0,95	0,2
31 et +	10	200	10	0,05	1	0,05
	200			1		

* 80 % des livraisons comportent moins 20 colis.
* 45 % des livraisons comportent plus de 16 colis.

III- Typologie des tableaux :

Selon le nombre de variables observées sur une même unité, plusieurs tableaux sont possibles :
* Tableau à simple entrée c'est à dire qui étudie une seule variable ;
* Tableau à double entrée c'est à dire qui étudie deux variables.

☛ **Exemple :**

Statistique du personnel d'une entreprise en fonction des salaires (x_i) et de l'âge (y_i).

Chapitre 3 : GRAPHIQUES STATISTIQUES DE BASE

La représentation graphique des fréquences simples ou cumulées relatives à une variable statistique donne lieu aux distinctions entre graphiques de distribution et graphiques de répartition.

I- Graphiques de fréquences simples, fonction de distribution :

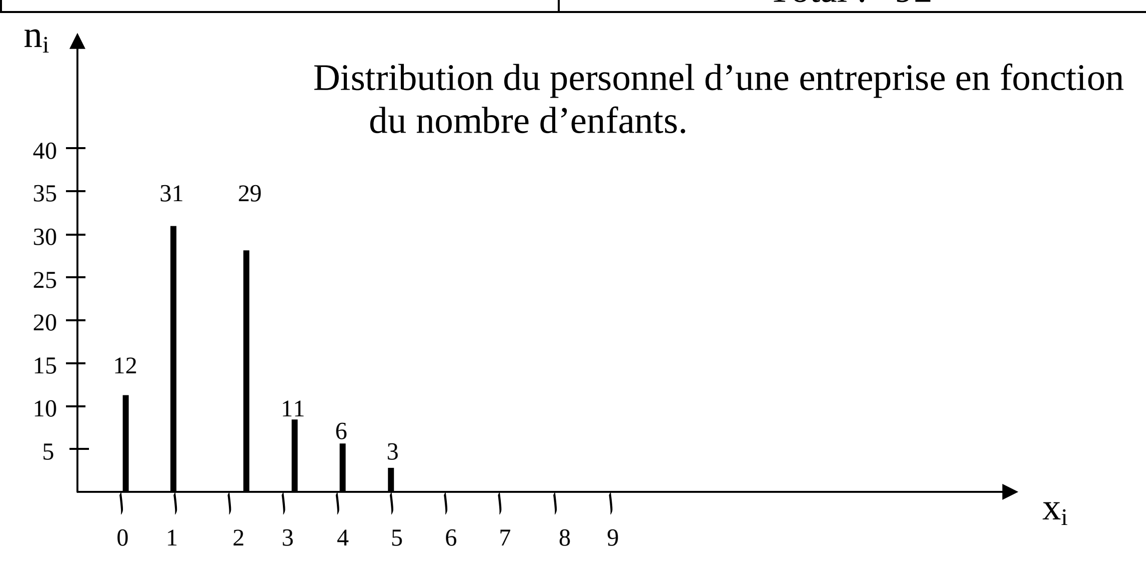
1- Cas d'une variable discrète ou discontinue :

La représentation graphique des fréquences simples d'une variable discrète peut s'effectuer sous la forme de graphique en bâton.
La valeur observée de la variable est portée sur l'axe des abscisses de la variable, et la fréquence simple correspondante est portée sur l'axe des ordonnés, cette dernière peut être exprimer en valeur relative ou en valeur absolue, le bâton est un segment de droite perpendiculaire à l'axe des abscisses (ox), de langueur ou de hauteur proportionnelle à l'effectif correspondant.

☛ **Exemple :**

Distinction du personnel d'une entreprise en fonction du nombre d'enfants.

Nombre d'enfants par personne (x_i)	Le nombre de personnes concernées (n_i)
0	12
1	31
2	29
3	11
4	6
5	3
	Total : 92



☛ **Remarque :**

Dans le cas d'une variable discrète, il ne faut pas joindre les sommets des bâtons car par définition, il n'existe pas de valeurs intermédiaires entre deux positions de la variable.

2- Cas d'une variable statistique continue :

La représentation graphique des fréquences simples d'une variable continue peut s'effectuer sous la forme d'un histogramme.
En portant dans l'axe des abscisses les valeurs des classes du caractère, et en leur donnant les fréquences correspondantes, ou représente la structure de la population étudiée.

a- Principes de construction de l'histogramme :

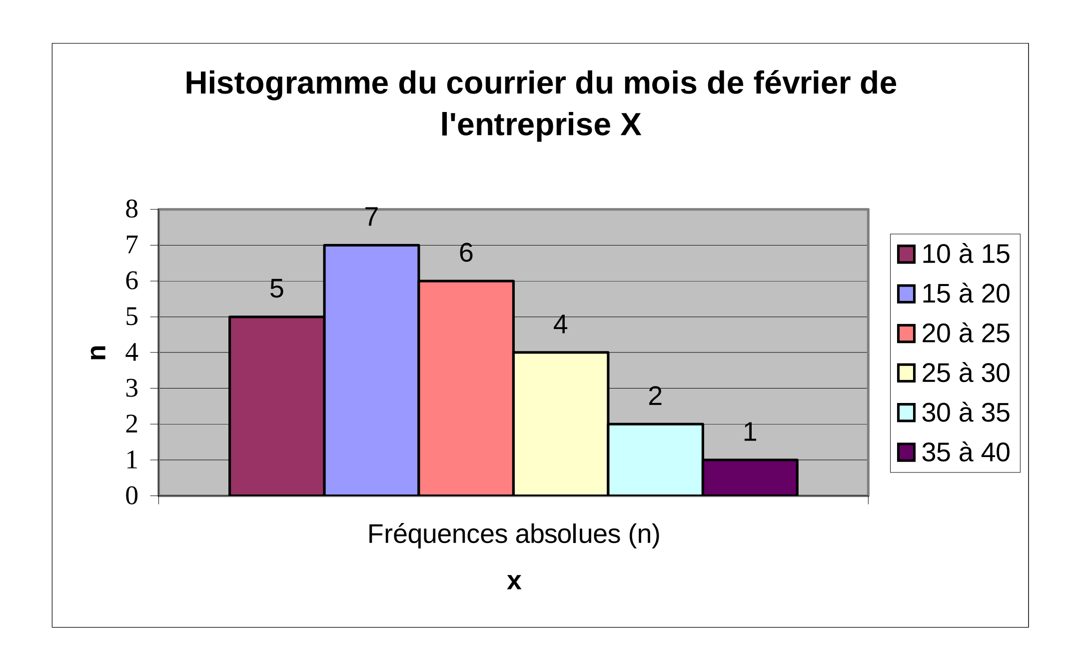
Pour chaque classe, on élève un rectangle ayant une base proportionnelle à l'intervalle de classes, et hauteur proportionnelle à la fréquence simple. Dans ce cas, ce sont les surfaces et non les hauteurs qui sont proportionnelles à l'effectif.
Dans la pratique deux cas peuvent se présenter :

☛ **Cas d'amplitude égale :**

- **Exemple :**

Distribution du courrier du mois de février d'une entreprise en fonction du nombre des lettres reçues.

Nombre des lettres reçues (x_i)	Fréquences absolues (n_i)	Amplitudes
10 à 15	5	5
15 à 20	7	5
20 à 25	6	5
25 à 30	4	5
30 à 35	2	5
35 à 40	1	5
	Total : 25	



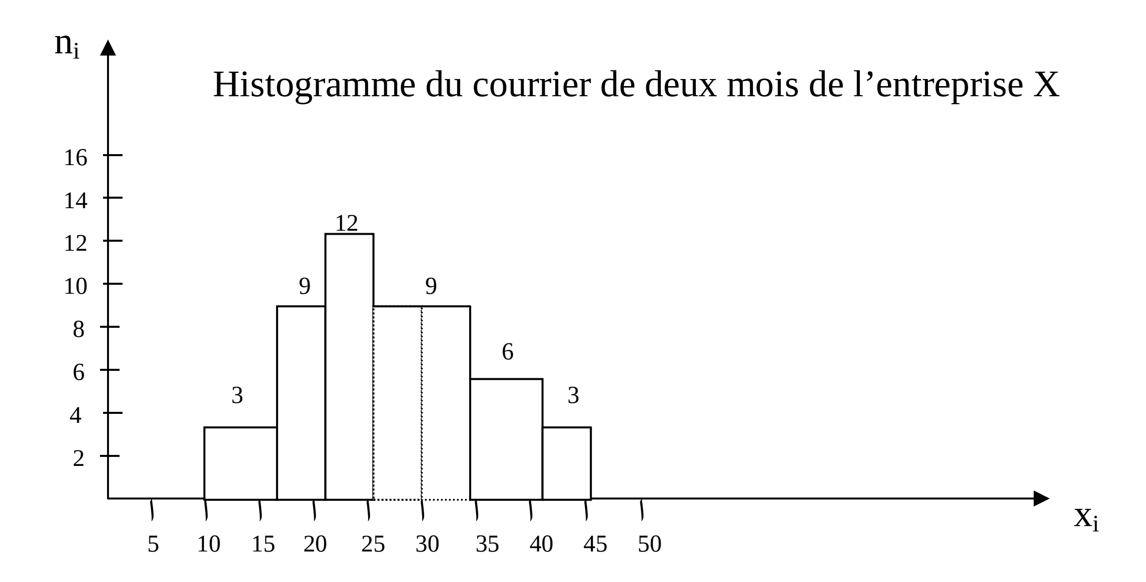
L'histogramme est constitué par l'ensemble des rectangles adjacent. Nous vérifions par exemple le rectangle représentatif de la classe 30 à 35 (effectif 2) a une surface double de celle du rectangle représentatif de la classe 35 à 40 (effectif 1).

☛ Cas d'amplitude inégale :

Dans ce cas, on respecte la proportionnalité des surfaces, il faut rectifier en conséquence les hauteurs.
Supposant que la distribution précédente, sur une période de deux mois, se présente de la façon suivante :

Nombre des lettres reçues (xi)	Fréquences absolues (n_i)	Amplitude
10 à 15	3	5
15 à 20	9	5
20 à 25	12	5
25 à 35	18	10
35 à 40	6	5
40 à 45	3	5
	Total : 51	

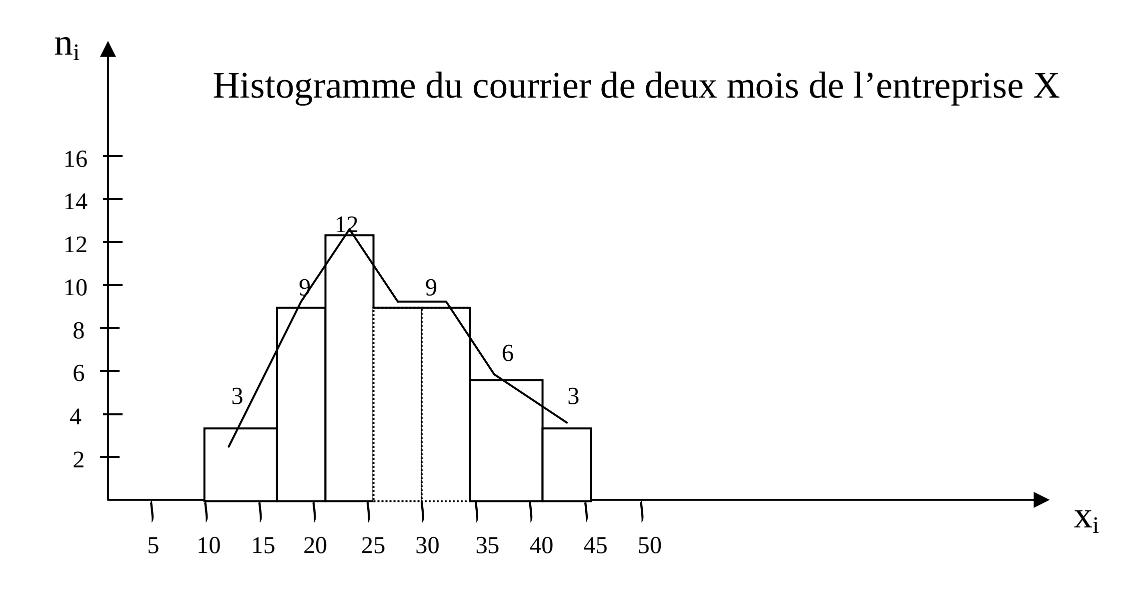
En effet, l'intervalle de la classe 25 à 35 est double des autres, ce qui se traduit si on ne modifie pas les fréquences par une marque de proportionnalité entre les surfaces.
Il convient donc de rendre les classes égales. L'amplitude de la classe 25 à 35 étant le double des amplitudes des autres classes, il faut avant toute représentation graphique de diviser par 2 la fréquence correspondante de la classe.



Le segment pointillé indique la transformation effectuée et met en garde le lecteur sur le caractère relatif des effectifs répartis entre les classes 25 à 30 et 30 à 35.

b- Le polygone des fréquences :

Le polygone des fréquences obtenues en joignant par des segments de droites au milieu des bases supérieurs des rectangles permet de rendre compte de la continuité de la variable.
En titre d'exemple, on peut prendre la figure précédente :



II- Graphique de fréquences cumulées, fonction de répartition :

1- Cas d'une variable statistique directe :

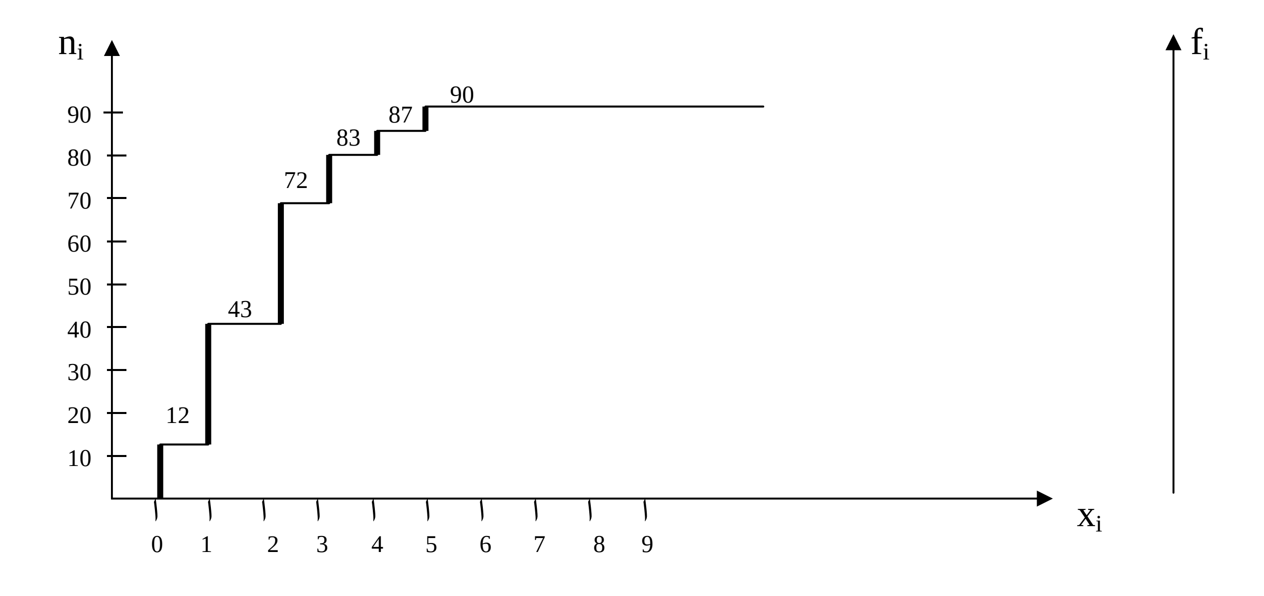
La représentation graphique des fréquences cumulées d'une variable directe s'effectue sous la forme de graphique en escalier.
« Les sauts » correspondent aux valeurs possibles de la variable et ils sont égaux aux fréquences cumulées croissantes ou décroissantes.

☛ **Exemple :**

Répartition du personnel d'une entreprise en fonction du nombre d'enfants.

x_i	Fréquences absolues			Fréquences relatives		
	Simples	Cumulées		Simples	Cumulées	
		Croissantes	Décroissantes		Croissantes	Décroissantes
0	12	12	89	0,135	0,135	1
1	31	43	77	0,348	0,483	0,865
2	29	72	46	0,326	0,804	0,517
3	11	83	17	0,124	0,936	0,191
4	4	87	6	0,045	0,977	0,067
5	2	89	2	0,022	1	0,022
	89			1		

- **Répartition croissante :**

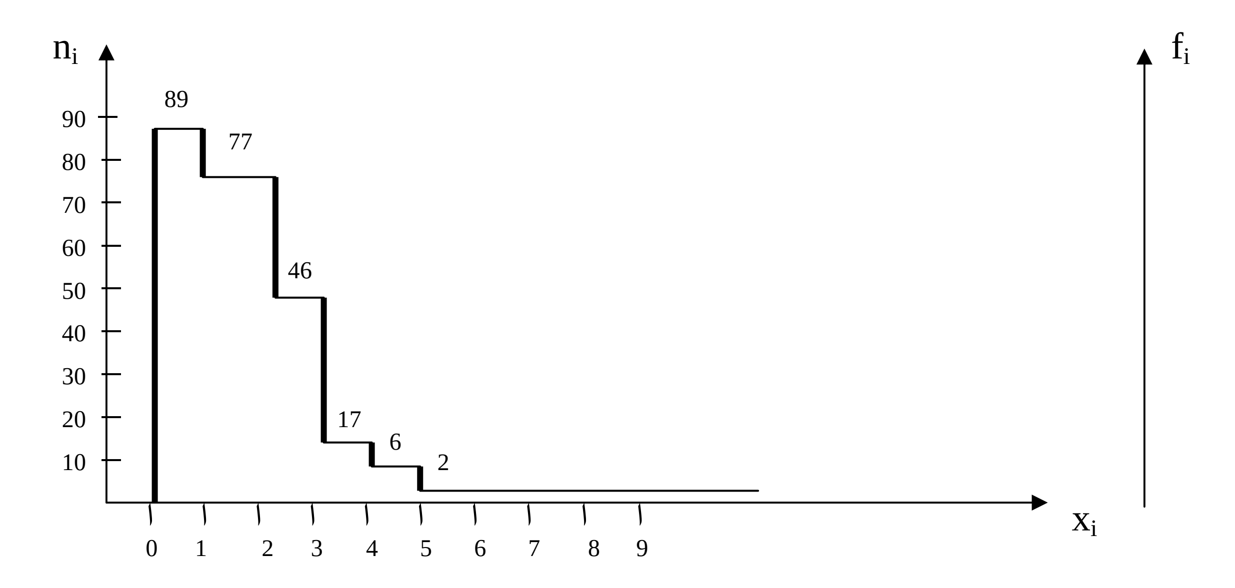


La lecture de ce graphique permet d'identifier la répartition (absolue ou relative) du personnel ayant x enfants ou moins.

- Exemple :

43 salariés 48% des salariés ont un enfant ou moins.

- **Répartition décroissante :**



Ici les résultats donnés sont inversés, à savoir x enfants ou plus.

- Exemple :

77 salariés ou 86% des salariés ont un enfant ou plus.

2- Cas d'une variable statistique continue :

La courbe des fréquences cumulées croissantes se construit en portant les points correspondants à chaque classe à la limite supérieure de l'intervalle de classes.

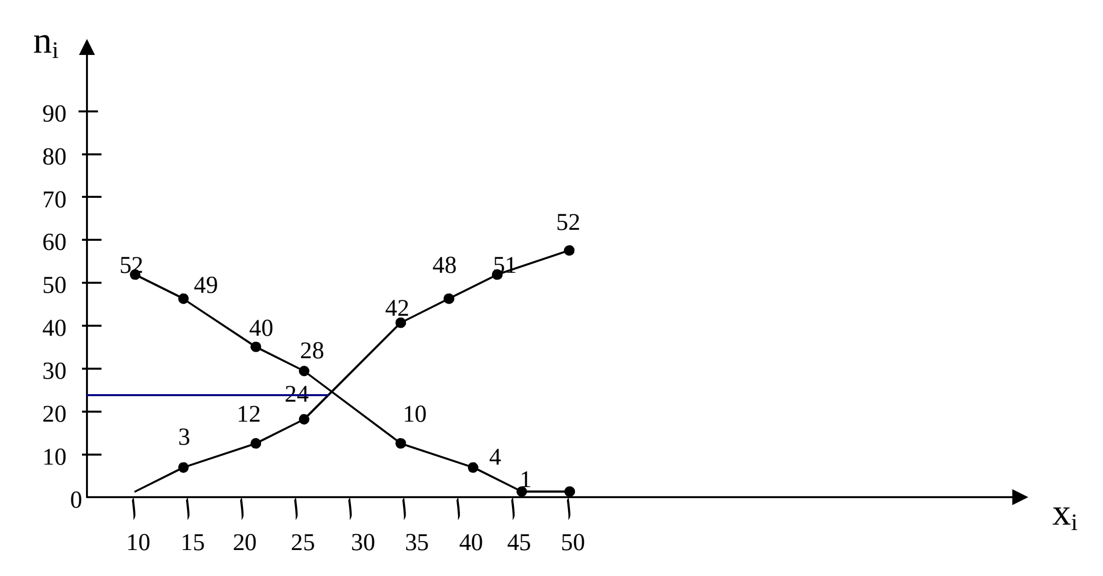
La présence de classes d'amplitude inégale n'entraîne aucune modification en ce qui concerne la construction de cette courbe.

La courbe des fréquences cumulées décroissantes se construit en portant les points correspondants à chaque classe à la limite inférieure de l'intervalle de classes.

☛ **Exemple :**

Soit le tableau suivant :

x_i	Fréquences absolues			Fréquences relatives		
	Simples	Cumulées		Simples	Cumulées	
		Croissantes	Décroissantes		Croissantes	Décroissantes
10 à 15	3	3	52	0,057	0,057	1
15 à 20	9	12	49	0,173	0,230	0,943
20 à 25	12	24	40	0,230	0,460	0,770
25 à 35	18	42	28	0,346	0,806	0,540
35 à 40	6	48	10	0,115	0,921	0,194
40 à 45	3	51	4	0,057	0,978	0,074
45 à 50	1	52	1	0,019	1	0,022
	52			1		



On contrôle l'exactitude du graphique en vérifiant que l'intersection des deux courbes a pour abscisse la moitié de l'effectif.

Chapitre 4 : LES CARACTERUSTIQUES DE TENDANCE CENTRALE ET DE POSITION.

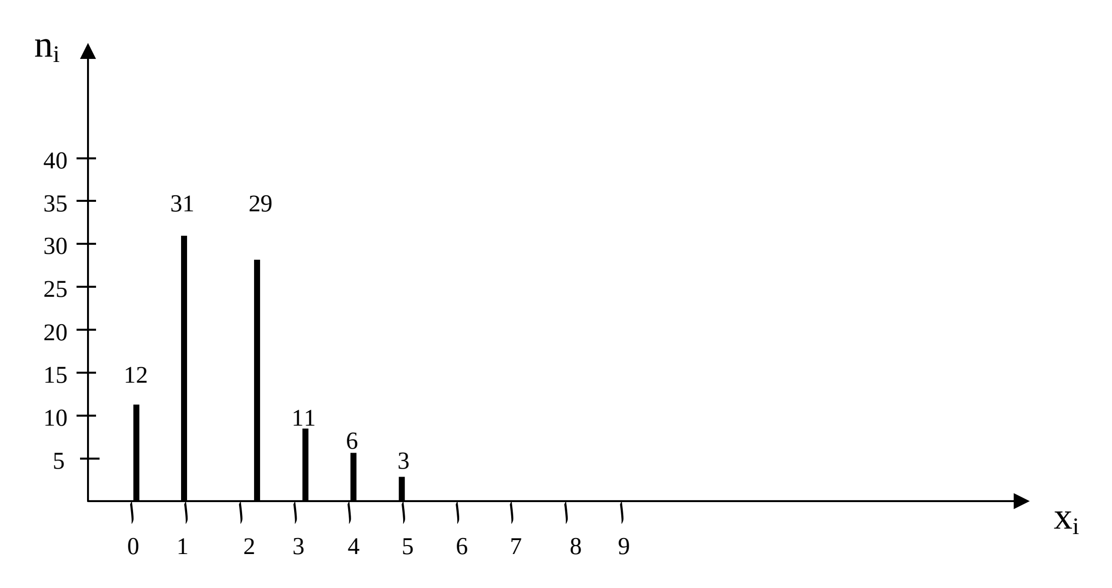
I- Le mode :

1- Définition :

Le mode est la valeur de la variable qui correspond à la fréquence maximale.
Dans le cas d'une variable discontinue, le mode se détermine de la façon suivante :

☛ **Exemple :**

x_i	n_i
0	12
1	31
→ **2**	**29**
3	11
4	6
5	3
	Total : 92



Donc le mode est 25, qui correspond au battons le plus élevé du graphique.

Dans le cas d'une variable continue, le mode s'applique à la classe qui correspond à la fréquence maximale. Celle ci s'appelle « classe modale ».

☛ **Remarque :**

La valeur du mode dans ces conditions dépend de l'amplitude des classes et qu'il faut vérifier l'égalité des intervalles de classes.

☛ **Exemple :**

x_i	n_i	Amplitude	Effectifs corrigés
100 à 110	8	10	8
110 à 120	22	10	22
120 à 125	18	5	36
125 à 130	**20**	**5**	**40**
130 à 140	12	10	12
140 à 160	6	20	3

Dans ce cas, la classe modale est « 125 à 130 ».
Le mode = 125 + (4/ 4+28). 5 = 125,625

En général, la formule du mode est la suivante :

$$\text{Mode (Mo)} = l_1 + \frac{d_1}{d_1 + d_2} \cdot a$$

II- La médiane :

1- Définition :

La médiane (M) est la valeur de la variable qui partage l'effectif en deux parties égales, les éléments de la population étant rongés par ordre croissant ou décroissant. En d'autre terme, la médiane est la valeur statistique qui correspond sur la courbe cumulative à une ordonnée représentant une fréquence relative 0,5 ou 50%, ce qui entraîne que 50% des observations seront inférieures à la médiane et 50% seront supérieures à la médiane.

a- Cas d'une variable discrète ou discontinue :

Il n'existe pas, en général, de valeur médiane, sauf dans l'hypothèse où la série possède un nombre de terme impair est connu individuellement.

☛ **Exemple :**

30 4 8 10 6 12 13 15 16

4 6 8 10 12 13 15 16 30

4 termes ⇓ 4 termes

Médiane

b- Cas d'une variable continue :

La valeur de la médiane peut de déterminer soit par calcul numérique, soit graphiquement.

☛ **Par le calcul numérique :**

L'équation F (M) = 0,5 admet une solution dépendante de l'effectif étudié :
* Partager le total des effectifs en deux.
- Si le total des fréquences simples où effectif est impaire, c'est à dire (2n+1), le rang de la médiane va être égale à [(2n+1) +1] / 2.
- Si le total des fréquences simples où effectif est paire, c'est à dire 2n/2 et (2n+1) / 2.

* Rechercher la classe correspondante aux rangs déterminés ;
* Déterminer au postulant une répartition homogène des valeurs dans la classe la valeur de la médiane par interpolation linéaire.

☛ **Exemple :**

X_i	Fréquences		
	Simples	Cumulés	
		Croissantes	Décroissantes
1000 à 1500	6	6	65
1500 à 2000	12	18	59
2000 à 2500	25	43	47
2500 à 3000	17	60	22
3000 à 3500	5	65	5
	65		

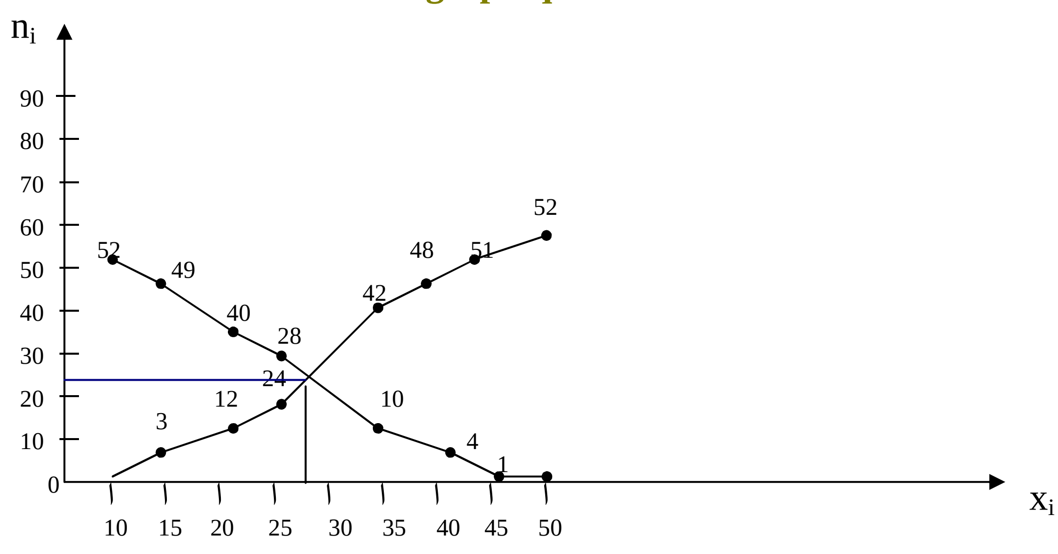
* Le rang est le suivant :
[(2n+1) + 1] / 2 = 33ème rang.

* Recherche de la classe correspondante au 33ème rang :
Nous savons d'après le tableau que la valeur de 2000 correspond au 18ème rang, et que la valeur de 2500 correspond au 43ème rang. La valeur du 33ème rang se trouve comprise entre 2000 et 2500, la classe médiane est donc 2000 à 2500.
La valeur exacte de la médiane est alors déterminée par l'application d'une règle de trois :
Médiane = 2000 + x.
Calculant cette valeur de x :
Sur l'intervalle de cet effectif occupe donc un sous intervalle de 500/25 = 20.
Pour arriver au 33ème rang, il faut donc ajouter 33-18 = 15 sous intervalles.
Donc x= 20×15 = 300.
Et puis la médiane = 2000+300 = 2300.

Autre solution plus simple :

M = 2000 + 500 [(33-18) / (43-18)].
M = 2300.

2- Détermination graphique :



Si on prend en titre d'exemple ce graphique, la médiane dans ce cas est d'environ 27.

☛ **Exercice :**

On donne la répartition de 100 ouvriers d'une entreprise selon leur salaire journalier.

Salaires journaliers	n_i	Fréquences cumulées croissantes	Fréquences cumulées décroissantes
80 - 120	10	10	100
120 - 160	30	40	90
160 - 200	40	80	60
200 - 240	20	100	20
	100		

Médiane = 160 + 40 [(50-40) (80-40)] = 170.
Donc50% des ouvriers touches un salaire journalier plus que 170 DHS. Et 50% des salariés touchent moins de 170 DHS.

Chapitre 5 : LES MOYENNES

Une moyenne paramètre de tendance centrale, est un nombre dont la détermination utilise l'ensemble des valeurs de la variable.
En règle générale, une moyenne est définie de la façon suivante :

$\overline{X} = 1/n\ [f(x_1) + f(x_2) + f(x_3) + \ldots + f\ (x_n)]$

La fonction f(x) étant une fonction continue monotone, croissante ou décroissante.
Selon l'expression de f(x), on peut distinguer la moyenne arithmétique, la moyenne géométrique, harmonique et quadratique.

I- La moyenne arithmétique :

Cette moyenne s'écrit en abrégé : $\overline{X}$

1- Définition :

a- Moyenne arithmétique simple :

Si la variable statistique est donnée sous forme de série (x_1, x_2, x_3 … x_n), la moyenne arithmétique est le rapport :

$$\overline{X} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n} = (1/n).\ \sum x_i$$

☛ **Exemple :**

Les notes obtenues par un candidat à un examen sont les suivants : 8 – 9 – 12 – 14 – 17.

$$\overline{X} = \frac{8 + 9 + 12 + 14 + 17}{5} = 12.$$

b- Moyenne arithmétique pondérée :

Lorsque la variable statistique est donnée sous forme de tableau de distribution, c'est-à-dire quand les valeurs de la variables sont affectées de fréquence, la moyenne arithmétique s'écrit de la façon suivante :

$$\overline{X} = \frac{\sum n_i\ x_i}{\sum n_i} = 550/50 = 11.$$

☛ **Exemple :**

Notes (x_i)	Effectifs (n_i)	$n_i\,x_i$	f_i	$f_i\,x_i$
4	1	4	0,02	0,08
6	2	12	0,04	0,24
8	4	32	0,08	0,64
9	7	63	0,14	1,26
10	10	100	0,2	2
12	13	156	0,26	3,12
13	5	65	0,1	1,3
14	6	84	0,12	1,68
17	2	34	0,04	0,68
	50	550	1	11

$$\overline{X} = \frac{\sum f_i\,x_i}{\sum f_i} = 11/1 = 11.$$

Dans le cas de variables groupées en classes, le calcule est le même en prenant pour valeur de la variable le centre des classes.

2- Calcule en cas d'une variable groupée en classes :

a- Calcule par changement d'origine :

La formule de la moyenne est la suivante :

$$\overline{X} = x_0 + \frac{\sum n_i\,(x_i - x_0)}{\sum n_i}$$

☛ **Exemple :**

classes	n_i	x_i	$x_i - x_0$	$n_i\,(x_i - x_0)$
40 - 45	8	42,5	-15	-120
45 - 50	12	47,5	-10	-120
50 - 55	28	52,5	-5	-140
55 - 60	30	57,5	0	0
60 - 65	58	62,5	5	290
65 - 70	29	67,5	10	290
70 - 75	7	72,5	15	105
	172			305

x_i, c'est le centre des calasses qui est calculer comme suit : (40 + 45) / 2

Donc :

$$\overline{X} = x_0 + \frac{\sum n_i (x_i - x_0)}{\sum n_i} = 59{,}27.$$

Avec : $x_0 = 57{,}5$.

b- Calcule par changement d'échelle :

Si les valeurs de la série sont des multiples entiers d'une valeur « k » (qui peut être également à l'intervalle de classes). Il est recommandé de prendre « k » comme unité.

$$\overline{X} = x_0 + k \frac{\sum n_i [(x_i - x_0) / k]}{\sum n_i}$$

☛ **Exemple :**

classes	n_i	x_i	$(x_i - x_0)/k$	$n_i [(x_i - x_0) /k]$
40 - 45	8	42,5	- 3	- 24
45 - 50	12	47,5	- 2	- 24
50 - 55	28	52,5	- 1	- 28
55 - 60	30	57,5	0	0
60 - 65	58	62,5	1	58
65 - 70	29	67,5	2	58
70 - 75	7	72,5	3	21
	172			61

$$\overline{X} = x_0 + k \frac{\sum n_i [(x_i - x_0) / k]}{\sum n_i} = 59{,}27.$$

Avec : « k » = l'amplitude.

II- Moyenne géométrique :

Lorsque les valeurs d'une série statistique varient en gros selon une progression géométrique, il est préférable de substituer à la moyenne arithmétique la moyenne géométrique (x_G).

1- Définition :

La moyenne géométrique de « n » valeur positive d'un caractère est la racine énième du produit de ces valeurs.

$$x_G = \sqrt[n]{x1.x2.x3....xn}$$

☛ **Exemple:**

Le chiffre d'affaire mensuel d'un nouveau produit a été au cours des six derniers mois : 256 – 332 – 432 – 562 – 731 – 950 .

$$x_G = \sqrt[6]{256*332*432*562*731*950} = 492.$$

☛ **Généralisation:**

Lorsque les valeurs de la variable sont affectées d'une fréquence, la moyenne géométrique est donnée par la formule suivante :

$\text{Log } x_G = 1/n\ (\Sigma\ n_i\ .\ \log x_i)$.

☛ **Exemple :**

x_i	n_i	$\log x_i$	$n_i\ .\ \log x_i$
2	4	0,301	1,204
4	5	0,602	3,01
8	8	0,403	7,224
16	2	1,204	2,408
12	1	1,505	1,505
	20		15,351

En appliquant la dernière formule :

$\text{Log } x_G = 0{,}76755$

$\Rightarrow x_G = 5{,}8.$

III- Moyenne harmonique :

1- Définition :

La moyenne harmonique est l'inverse des moyennes arithmétiques des inverses.

$$x_H = \frac{\sum n_i}{\sum n_i \times 1/n}$$

IV- Moyenne quadratique :

1- Définition :

La moyenne quadratique est définie par la formule suivante :

$$x_Q = \sqrt{\frac{\sum n_i x_i^2}{\sum n_i}}$$

2- Observation :

Il faut savoir que :

$x_Q \geq \bar{x} \geq x_G \geq x_H.$

Chapitre 6 : LES PARAMETRES DE DISPERSION.

I- L’étendu :

Ce paramètre également appelé « intervalle de variation ».

1- Définition :

L’étendu d’une série statistique est la différence entre la plus grande valeur et la plus petite valeur du caractère.

$e = x_M - x_m$.

2- Application :

Soit la série statistique suivante :

x_i	n_i
100	2
105	15
110	28
115	16
120	3

e = 120 – 100 = 20.

3- Commentaire :

La simplicité de ce calcule ne doit pas le faire oublier que l’étude est très sensible aux fluctuations des valeurs (valeurs extrêmes) qui sont souvent peu représentatives. Cette valeur caractéristique qui correspond à un concept, forme utilisée dans la pratique. (L’écart entre le premier et le dernier coureur, l’écart entre la meilleure et plus faible note).

II- Les intervalles :

Ces caractéristiques permettent une mesure de dispersion qui élimine l’influence des valeurs extrêmes.

1- Les intervalles interquartiles :

a- Définition :

L’intervalle interquartile d’une série statistique est égal à la différence :

$Q_3 - Q_1$.

x_i	n_i	$n_i \uparrow$
1000 - 1500	6	6
1500 - 2000	8	14
2000 - 2500	3	17
2500 - 3000	1	18
	18	

* Calculer Q_1 :
Rang = 18/4 = 4,5.
$Q_1 = 1000 + 500\ [(4{,}5 - 0) / (6 - 0)]$.
$Q_1 = 1375$.

* Calculer Q_3 :
Rang = (18/4).3 = 13,5.
$Q_3 = 1500 + 500\ [(13{,}5 - 6) / (14 - 6)]$.
$Q_3 = 1968{,}75$.

$Q_3 - Q_1 = 1968{,}756 - 1375 = 593{,}75$.

On trouve dans cet intervalle 50% des observations concentrées autour de la médiane. Plus l'intervalle est réduit, plus la concentration autour des valeurs centrales est forte.

III- La variance, l'écart type :

1- Définitions :

* La variance d'une série statistique est la moyenne arithmétique des carrés des écarts

$$V = \frac{\sum [n_i (x_i - \bar{x})^2]}{\sum n_i}$$

* L'écart type est la racine carrée de la variance :

$$\gamma = \sqrt{V}$$

2- Application:

a- détermination directe de la variance :

x_i	n_i	$n_i\ x_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$n_i\,(x_i - \overline{x})$	$n_i\ x_i^2$
17	1	17	- 7	49	49	289
19	1	19	- 5	25	25	361
21	2	42	- 3	9	18	882
23	6	138	- 1	1	6	3174
25	5	125	1	1	5	3125
27	3	81	3	9	27	2187
29	2	58	5	25	50	1682
31	0	0	7	49	0	0
	20	480			180	11700

$$\overline{X} = \frac{\sum f_i\ x_i}{\sum f_i} = 24.$$

$$\gamma = \sqrt{V} = 3.$$

Lorsqu'une moyenne arithmétique est une valeur entière, les calcules sont assez simples, mais la plupart du temps la moyenne est un nombre décimal, ce qui rend l'élévation au carrée des écart plus difficile. L'hypothèse de non utilisation de machines, nous proposons les améliorations du calcul suivantes :

b- Autres modes de détermination :

- **Expression développée de la variance :**

$$V = \frac{\sum (n_i\ x_i\ )^2 - \ \overline{x}^2}{\sum n_i}$$

- **Calcul de la variance, formule simplifiée :**

Comme pour la moyenne, le recours aux procédés de simplification peut être utile.

*** Calcul de variance par changement d'origine :**

En posant une valeur arbitraire x_0 (la plupart du temps cette valeur sera centrale) l'expression de la variance devient :

$$V = \frac{\sum n_i \ (x_i - x_0)^2 - (\overline{x} - x_0)^2}{\sum n_i}$$

*** Calcul de la variance en faisant un changement d'échelle :**

Dans les séries à intervalles de classes égaux, il est conseillé de trouver un nombre « k » qui signifie en maximum l'écart $(x_i - x_0)$. On peut alors écrire :

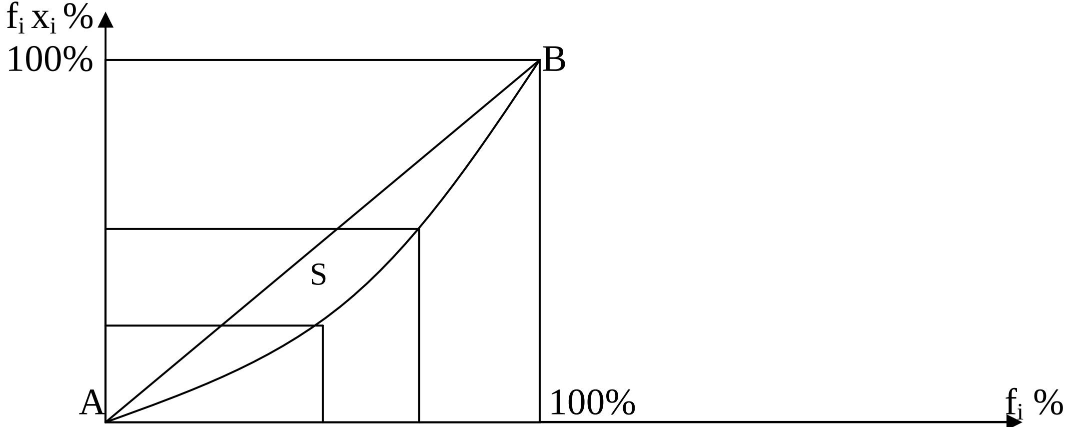
$$V = \frac{k^2 . \sum n_i \ [(x_i - \overline{x}) / k]^2}{\sum n_i}$$

La combinaison des deux procédés de simplification permet d'écrire la formule générale :

$$V = \frac{k^2 . \sum n_i \ [(x_i - x_0) / k]^2 - (\overline{x} - x_0)^2}{\sum n_i}$$

Chapitre 7 : LA CONCENTRATION.

La notion de concentration est apparentée à celle de dispersion. On sait l'importance tenue par l'idée de concentration dans les phénomènes économiques.
Soit un ensemble statistique dont chaque élément est affecté d'un caractère susceptible d'addition. Ainsi un ensemble d'individus ou d'unité de production classés selon la fortune, le revenu, les salaires versés ou les salariés occupés, le chiffre d'affaire, et enfin la surface exploitée.
Il est possible de classer ces unités :
* Selon leur nombre ;
* Selon l'importance du caractère procédé.
Cette distinction conduis à une double figuration, c'est-à-dire à conduire :
* Un histogramme donnant le nombre des effectifs par classes ;
* Un histogramme donnant l'importance du caractère procédé par classes.
On peut à partir de ces histogrammes déterminer mes médianes.
Dans l'histogramme donnant l'importance du caractère procédé par classes, la médiane porte un nom spécial « médiale ».
La médiale est une valeur particulière du caractère. Cette valeur du caractère est tel que tous les caractères supérieurs, constitue une moitié de la masse globale des caractères, l'autre moitié étant constituée par tous les caractères inférieurs.
L'analyse de la concentration pourrait résulter de la comparaison des deux histogrammes, plus précisément de la mesure de l'écart entre la médiale et la médiane.
On procède cependant autrement, en exprimant les deux distributions sur un même graphique. Ce graphique porte le nom de courbe de concentration ou courbe de GINI ou courbe de Lorenz.
Cette courbe se construit à partir des fréquences cumulées relatives en pourcentage.
Dans un graphique cartésien on porte :
* En abscisse, les fréquences relatives cumulées du nombre des effectifs (f_i) ;
* En ordonné, les fréquences relatives cumulées du caractère procédé par classes.
Le résultat de la concentration est une courbe obtenue point par point.



La concavité de la concentration est toujours dirigée vers le bas. La surface « S » s'appelle l'aire de la concentration. Quand l'aire est nulle, il y a absence de concentration, quand elle est égale à la surface du demis carrée, il y a concentration totale.
On définis ainsi un indice de concentration par le rapport suivant :

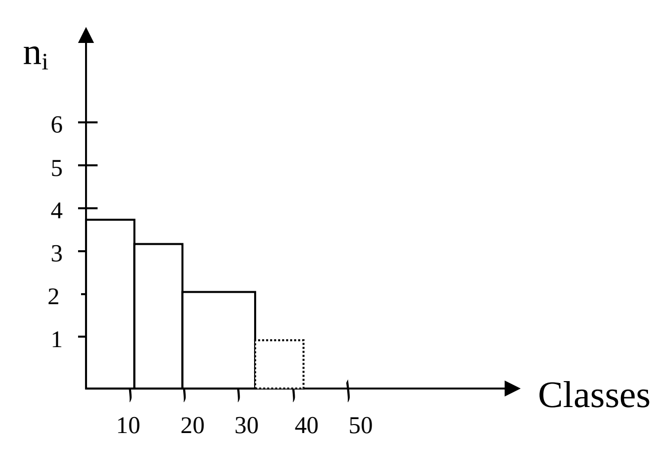
C = (5000 – S) / 5000. (Coefficient de concentration).
Cet indice varie entre 0 et 1.

☛ **Exemple :**

Soit la distribution statistique donnant le nombre de salariés par classes de salaire.

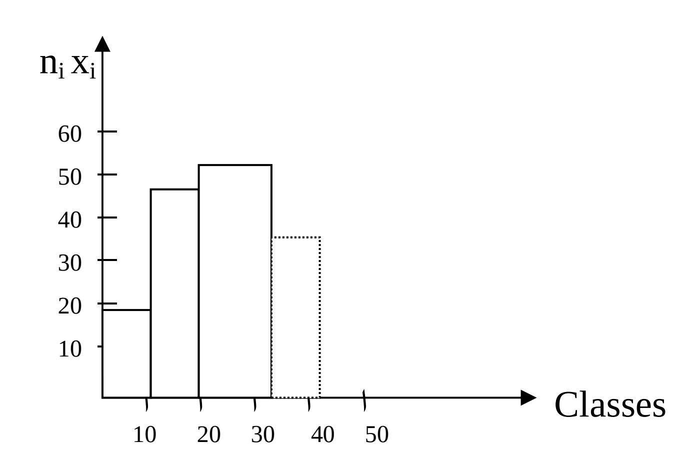
Classes de salaire	Nombre de salariés (n)	n_i ↑	x_i	$n_i x_i$	$n_i x_i$ ↑	f_i	f_i ↑ en %	$f_i x_i$	$F_i x_i$ ↑ en %
0 - 10	4	4	5	20	20	0,4	40	0,13	13
oct-20	3	7	15	45	65	0,3	70	0,3	43
20 - 30	2	9	25	50	115	0,2	90	0,33	76
30 - 40	1	10	35	35	150	0,1	100	0,23	100
	10			150					

* Construire l'histogramme donnant le nombre des effectifs par classes.



* Déterminer la médiane :
Rang : 10/2 = 5.
Médiane = 10 + 10[(5-4)/(7-4)] = 13,33.

* Construire l'histogramme donnant l'importance du caractère procédé par classes.

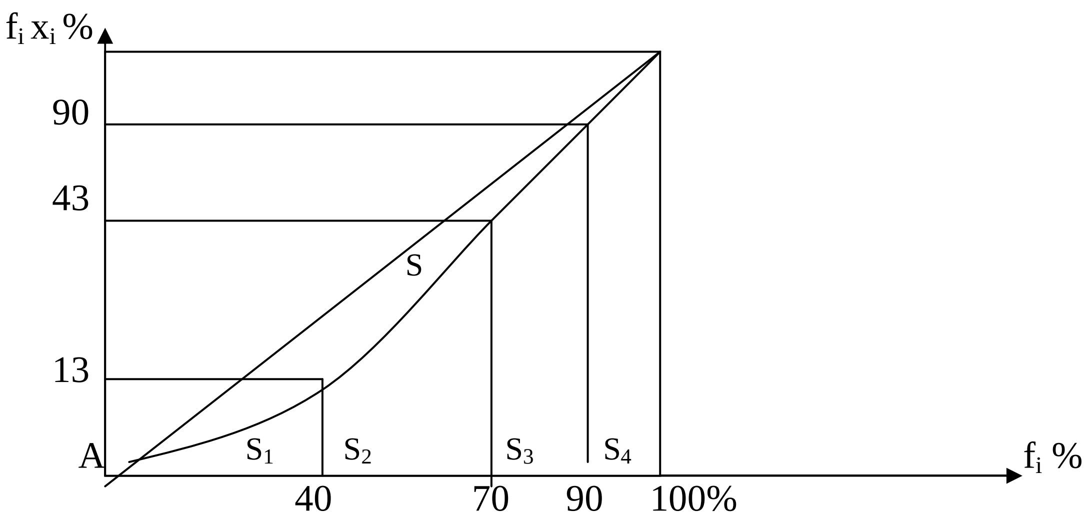


* Déterminer la médiale :
  Rang : 150/2 = 75.
  Médiale = 20 + 10 [(75 – 65) / (115 – 65)] = 22.

* Que peut on dire de la différence : Ml – Me = ΔM.
  22 – 13,33 = 8,66.

Si ΔM est grand par rapport au domaine de variation du caractère, la concentration est grande ;
Si ΔM est petit par rapport au domaine de variation du caractère, la concentration est petite.

* Construire la courbe de concentration :



$S_1 = (40 . 13) / 2 = 260.$
$S_2 = [(13 + 43).30] / 2 = 845.$
$S_3 = [(43 + 76).20] / 2 = 1190.$
$S_4 = [(76 + 100).10] / 2 = 880.$
$S_1 + S_2 + S_3 + S_4 = 3170.$

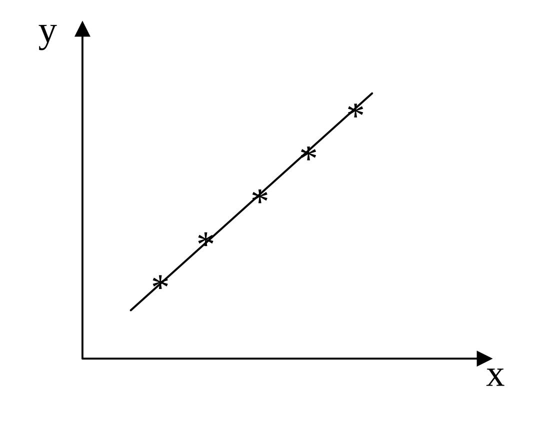
Donc : C = (5000 – 3170) / 5000 = 0,36.

Chapitre 8 : L'ETUDE DE LA CORRELATION.

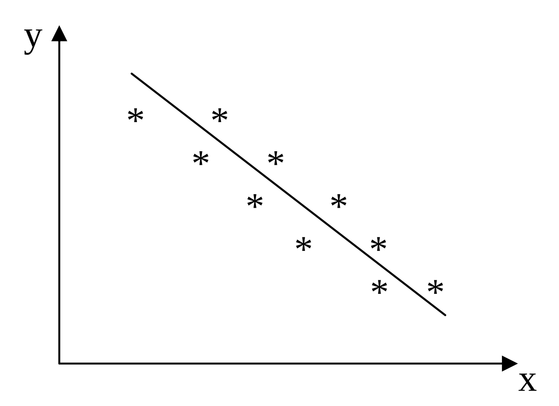
Afin de faciliter les prises de décisions, les responsables d'entreprises essayent de déterminer des indices annonciateurs du futur. Cette recherche de liaison entre les phénomènes peut être plus ou moins confié.
Par exemple, dans une entreprise, l'expérience montre que le chiffre d'affaire du printemps d'une année indique avec une certaine fiabilité celui de l'automne, et ceux malgré une absence apparente de cause entre les deux phénomènes.
Cependant, cette vision simpliste de relations n'indique pas ou mal la durée et l'intensité de ces liaisons. Il est donc nécessaire de les confier. La statistique permet de répondre à ce besoin en mesurant de la relation existante entre deux phénomènes, c'est la corrélation.
Enfin, l'existence de corrélation entre deux phénomènes n'implique pas obligatoirement une relation causale entre ces deux phénomènes.
Le chiffre d'affaire du printemps de l'année « n » ne présente pas la cause du chiffre d'affaire de l'automne suivant, il n'en est seulement qu'un indicateur.
Le lieu de corrélation entre deux phénomènes est un lien intermédiaire :
* La liaison fonctionnelle qu'on la note $y = f(x)$. Par exemple la circonférence d'un cercle « y » est fonctionnelle de la grandeur de son rayon « x ».
* L'indépendance totale. Par exemple l'évolution du prix de l'essence et celle des cotisations de la sécurité sociale.
C'est ce qui explique que la méthode de la corrélation se ramène au calcul de liaison fonctionnelle à une approximation pré.
Ca sera la démarche de notre développement :
* Les droites de régression ;
* Le coefficient de corrélation.

I- Les nuages de points :

Il s'agit des représentations graphiques des différentes courbes de deux caractères. Il permet de visualiser globalement les liens de dépendance statistique. Ce dernier, quand il existe peut être linéaire ou pas. Pour la simplicité de notre exposé, nous ne traitons que la corrélation linéaire.



Dépendance linéaire parfaite.



Dépendance linéaire forte.



II- Les droites de régression :

1- Remarque :

L'existence d'une relation linéaire ne suffit pas comme lien de cause à effet.
La mesure de la corrélation est purement mathématique et peut être effectué entre des phénomènes indépendants, il faudra donc toujours expliquer le pourquoi d'une forte corrélation.
Dans le cas d'une série de deux variables « x » et « y », il est possible de considérer successivement chaque variable comme variable expliquée, puis comme variable explicable. Dans ces conditions, nous pouvons calculer deux droites de régression.

a- La droite de régression de « y » en « x » d'équation y = ax + b :

Permettant de déterminer « y » connaissant « x ».

b- La droite de régression de « x » en « y » d'équation x = a'y + b' :

Permettant de déterminer « x » connaissant « y ».

☛ **Application :**

Une entreprise souhaite expliquer et prévoir ses ventes « y » par rapport à des dépenses de publicité engagées « x », ou au contraire déterminer les dépenses de publicité « y » en fonction de ces ventes « x ».
Pour concrétiser ces notions, nous allons utiliser l'exemple suivant :

Dépenses de publicité "X"	Vente "Y"	$X_i = x_i - \overline{x}$	$Y_i = y_i - \overline{y}$	$x_i y_i$	$(x_i - \overline{x})^2$	$(y_i - \overline{y})^2$
800	1500	-110	-700	77000	12100	490000
870	1900	-40	-300	12000	1600	90000
900	2000	-10	-200	2000	100	40000
920	2300	10	100	1000	100	90000
970	2500	60	300	18000	3600	90000
1000	3000	90	800	72000	8100	640000
5460	13200			182000	25600	1360000

1- Droite de régression de « y » en « x » :

* Choix des variables :
Nous somme dans le situation suivante : l'entrepreneur désir prévoir ses ventes (Y : variable expliquée) par rapport à des dépenses de publicité engagées (X : variable explicative).

* Caractéristique de la droite y = ax + b :
L'équation de cette droite se détermine en appliquant la méthode des moindres carrée.
Cette droite passe par le point moyen $(\bar{x}, \bar{y})$ du nuage des points et que la valeur de sa pente (a) se détermine par la formule suivante :

$$a = \frac{\sum (X_i Y_i)}{\sum X_i^2}$$

$\bar{x} = 5460/6 = 910.$ $\bar{y} = 13200 / 6 = 2200.$

$X_i = x_i - \bar{x}$ $Y_i = y_i - \bar{y}.$

Donc : b = 2200- (7,10 * 910) = - 4270.
a = 7,1.
Enfin : y = 7,1 x – 4270.

2- Droite de régression de « x » en « y » :

* Choix des variables :
Nous sommes dans la situation suivante : l'entrepreneur veut déterminer ses dépenses de publicité (qui devient variable expliquée « y ») en fonction de ses ventes qui devient variable explicative « X ».
* Caractéristique de la droite :
L'équation de cette droite se détermine de la façon suivante : elle passe par le point moyen(x , y) et a pour pente la formule suivante :

$$a' = \frac{\sum (X_i Y_i)}{\sum Y_i^2} = 182000/1360000 = 0,13.$$

$b' = \bar{x} - a'y' = 617,40.$

Donc: x = 0,13 + 617,40.

☛ **Commentaire:**

Il existe donc une liaison certaine entre les deux phénomènes. Certes les dépenses de publicité expliquent correctement les ventes, mais cette dernière influence certainement les dépenses de publicités futures qui à leur tour conditionnent les ventes.
En généralisant :
* Droite de régression de « y » en « x » : D (yx) ;
* Droite de régression de « x » en « y » : D' (xy) ;
* Graphique ;
* Conséquences.

a- Le dénominateur de « a » est la carré de l'écart type de la série de x_i (soit sa variance). De même pour « a' » le dénominateur représente la variance de la série « y ».
b- Les deux droites de régression ont des coefficients directeurs « a » et « a' » de même signe. En effet, les dénominateurs de celui-ci sont toujours positifs et leurs numérateurs sont identiques.
c- Les deux droites de régression « D » et « D' » ne sont confondues que dans le cas suivent : a = 1/a' $\Leftrightarrow$ a.a' = 1.
d- Le numérateur de « a » et de « a' » sont égaux, leur valeur commune $\sum(x_i y_i)$ s'appelle covariance de la série statistique.

III- Le coefficient de corrélation :

1- Définition :

Le coefficient de corrélation « r » est un indicateur de dépendance entre deux phénomènes. Ce concept est très utile dans la gestion et l'administration des entreprises. Il permet d'entrevoir, puis de vérifier l'existence d'un bien entre des phénomènes tel que les salaires et les prix, l'absentéisme et taux de primes, les accidents du travail et les heures supplémentaires…etc.
De façon graphique, le coefficient de corrélation indique le plus ou mois grand de degré de rapprochement des deux droites de régression.
Il est défini comme étant égale à la racine carrée du produit de la pente des deux droites de régression.

$$(r^2 = aa') \Rightarrow |r| = \sqrt{aa'}.$$

$$r = \sum (X_i\ Y_i) = \sqrt{\sum (X_i^2). \sum (Y_i^2)}$$

☛ **Remarque:**

Le coefficient de corrélation :
* Est un nombre sans dimension entre 0 et ±1 ;
* il est toujours du signe de $\sum (X_i\ Y_i)$ qui peut être positif, nul ou négatif.

☛ **Interprétation de la corrélation et de la régression :**

* Lorsque les points du nuage ne sont pas alignés, le coefficient de corrélation (r) est, en valeur absolue, inférieur à 1. Donc -1 < r < 1. Les deux droites de régressions sont alors distinctes (aa' < 1).
* La fidélité de la représentation du nuage des points par les droites de régression est fonction de la valeur des coefficients de corrélation. Plus cette dernière en valeur absolue s'approche de « 1 », plus cette fidélité est importante.
* si « r » est proche de « 1 », les deux phénomènes sont en relation étroite et leur sens de variation est identique : à un accroissement de « x » correspond un accroissement de « y ». Comme par exemple l'évolution des salaires et des prix.



* Si « r » est proche « -1 », les deux phénomènes sont en relation étroite, et leur sens de variation est inverse. Autrement dit un accroissement de « x » correspond à une diminution de « y ».
* Si « r » est comprise entre « - 0,5 » et « 0 ,5 », il n'y a pas de véritable relation linéaire entre « x » et « y ». Cela peut provenir d'une indépendance ou d'une relation non linéaire entre les deux phénomènes « x » et « y ».
Le nuage de points est dans ce cas très indicatif.
En règle générale, la corrélation :

- Est bonne si $|r| \geq 0,8$.
- Est moyenne si $0,5 < |r| < 0,8$.
- Est mauvaise si $|r| < 0,5$.